SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Effect of 6p21 region on lung function is modified by smoking: a genome-wide interaction study

Boram Park[1], Jaehoon An[1], Wonji Kim[2], Hae Yeon Kang[3], Sang Baek Koh[4], Bermseok Oh[5], Keum Ji Jung[6], Sun Ha Jee[6], Woo Jin Kim[7], Michael H. Cho[8,9], Edwin K. Silverman[8,9], Taesung Park[2,10]✉ & Sungho Won[1,2,11]✉

Smoking is a major risk factor for chronic obstructive pulmonary disease (COPD); however, more than 25% of COPD patients are non-smokers, and gene-by-smoking interactions are expected to affect COPD onset. We aimed to identify the common genetic variants interacting with pack-years of smoking on FEV_1/FVC ratios in individuals with normal lung function. A genome-wide interaction study (GWIS) on FEV_1/FVC was performed for individuals with FEV_1/FVC ratio ≥ 70 in the Korea Associated Resource cohort data, and significant SNPs were validated using data from two other Korean cohorts. The GWIS revealed that rs10947231 and rs8192575 met genome-wide significant levels; For $H_0 : \beta_{SNP} = \beta_{SNP*pack-years} = 0$ vs $H_1 : not H_0$, the likelihood ratio (LR) test was conducted, and its *P* values, P_{LR}, for rs10947231 and rs8192575 were 2.23×10^{-12} and 1.18×10^{-8}, respectively. Interaction between rs8192575 and smoking is significantly replicated with two additional data ($P_{INT} = 0.0454$, 0.0131). Expression quantitative trait loci, topologically associated domains, and PrediXcan analyses revealed that rs8192575 is significantly associated with *AGER* expression. SNPs on the 6p21 region are associated with FEV_1/FVC, and the effect of smoking on FEV_1/FVC differs among the associated genotypes.

Chronic obstructive pulmonary disease (COPD) is a common respiratory disease with high worldwide morbidity and mortality[1], characterized by progressive airflow obstruction[2]. Although cigarette smoking is the major environmental risk factor for COPD, multiple factors can contribute to this disease, including air pollution, infection, and asthma[1,3]. However, sensitivity to smoking differs among individuals, and only a minority of smokers develop COPD[4,5], highlighting the potential importance of genetic architecture. Severe α_1-antitrypsin (AAT) deficiency is the best known genetic risk factor for COPD[6], and genome-wide association studies (GWAS) have identified multiple promising candidate genes for COPD, including *FAM13A*, *HTR4*, *RIN3*, *HHIP*, *ADAM19*, *CHRNA3/5*, *AGER*, and *EEFSEC*[6–8]. Estimated heritabilities in family aggregation studies are typically 30%[9]. Notwithstanding the importance of genetic architecture, AAT deficiency occurs in only 1 ~ 2% of COPD patients[5] and the pathological role of most COPD candidate genes is unknown[10]. Although COPD results from genetic and environmental factors,limited information is available regarding the genetic factors that actually contribute to COPD, and gene-environment interactions, except for AAT deficiency, have been difficult to identify[11].

[1]Department of Public Health Sciences, Seoul National University, Seoul, South Korea. [2]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, South Korea. [3]Department of Internal Medicine, Healthcare Research Institute, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, South Korea. [4]Department of Preventive Medicine, Yonsei University Wonju College of Medicine, Wonju, South Korea. [5]Department of Biochemistry and Molecular Biology, School of Medicine, Kyung Hee University, Seoul, South Korea. [6]Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, South Korea. [7]Department of Internal Medicine and Environmental Health Center, Kangwon National University Hospital, School of Medicine, Kangwon University, Chuncheon, South Korea. [8]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [9]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [10]Department of Statistics, Seoul National University, Seoul, South Korea. [11]Institute of Health and Environment, Seoul National University, Seoul, South Korea. ✉email: tspark@stats.snu.ac.kr; sunghow@gmail.com

Nevertheless, COPD is known to be strongly influenced by cigarette smoking and multiple genetic variants, and recent studies have reported gene-environment interactions. For instance, Aschard et al.[12] calculated the genetic risk scores for 26 genome-wide significant single nucleotide polymorphisms (SNPs), and reported significant interactions between genetic risk scores and smoking. In addition, both Hancock et al.[13] and Park et al.[14] hypothesized the existence of SNP-by-smoking interactions,the first performed a genome-wide interaction study on pulmonary function, modelling the primary effects of single SNPs and their interactions, while the latter reported *SOX9*-by-smoking interactions. However, there is heteroscedasticity in pulmonary function between smokers and non-smokers. Furthermore, smoking has a nonlinear effect on pulmonary function[15], and the heterogeneity of this association has complicated SNP-by-smoking interaction analyses, thus limiting the number of identified interactions.

COPD is a progressive diseases that can be prevented but is irreversible[16], so the most effective way to prevent COPD is primary prevention. Primary prevention involves two concepts: (1) to keep healthy individuals consistently healthy (health promotion), and (2) to prevent the onset or exacerbation of the diseases (disease prevention)[17]. Therefore, it is important to maintain healthy lung function, and we focus on the genetic and environmental factors in people with healthy lung function. In our previous reports[14], we analyzed whole individuals on FEV_1, and thus in this article, we aimed to identify genetic variants interacting with smoking, using spirometry measurements of FEV_1/FVC ratios from individuals with $FEV_1/FVC \geq 70$ in the Korea Associated Resource (KARE), as this measurement determines the presence of airflow limitation and obstructive lung diseases such as COPD. We assessed significant SNPs on our genome-wide interaction study (GWIS) using Gene-Environment of Interaction and phenotype (GENIE) and Atherosclerosis Risk of a Rural Area in Korean General population (ARIRANG) data. We focused on potential SNP-by-smoking interactions considering only the SNPs that reached overall genome-wide significance levels.

Materials and methods

The KARE cohort was used for the discovery analysis in GWIS; GENIE and ARIRANG cohorts were included in validation analysis. Only participants with FEV_1/FVC ratio ≥ 70 were included.

Discovery analysis using KARE. The KARE project was initiated in 2007 for a large-scale GWAS, and participants constituting the independent Ansan and Ansung cohorts were included in the Korean Genome Epidemiology Study (KoGES)[18]. KoGES involved longitudinal prospective studies on 5018 participants in Ansung and 5020 participants aged 40–60 years in the Ansan area. KARE genotype data were obtained using the Affymetrix Genome-Wide Human SNP array 5.0[18], and quality control analyses were performed,8172 participants underwent spirometry analysis, and their smoking history was recorded (Fig. 1). Smoking history was obtained from questionnaires, and pack-years of smoking was considered for analyzing SNP-by-smoking interactions. Among the 8172 participants, 7473 participants showed FEV_1/FVC ratio ≥ 70; 4768 were non-smokers, 1140 were former smokers, and 1565 were current smokers.

Validation analyses using GENIE and ARIRANG data. The GENIE cohort comprised 7999 participants agreeing to provide blood samples and to participate in genetic studies that had visited Seoul National University Gangnam Center in 2014.[19] Participants underwent genotype analysis using the Affymetrix Axiom KORV1.1–96 Array[20] and genotype quality control (QC) was performed. The 4413 participants with FEV_1/FVC ratio ≥ 70, age > 40 years, available spirometric data, and smoking history were included in the association analysis. Based on questionnaires, 2520 individuals were non-smokers, 1380 were former smokers, and 513 were current smokers. From this dataset, the 2520 non-smokers and 513 current smokers were included in the GWIS validation study.

ARIRANG is an ongoing study on cardiovascular and metabolic risk factors, and participants aged 40–70 years are part of the KoGES study in rural Wonju and Pyengchang[21]. The ARIRANG genotype data were obtained using the Affymetrix Genome-Wide Human SNP array 6.0, and genotype QC was performed. Spirometry data and smoking history were available for 513 participants with FEV_1/FVC ratio ≥ 70. Based on questionnaires, 369 individuals were non-smokers, 65 were former smokers, and 79 were current smokers. All participants were included in the GWIS.

Quality control. For the discovery GWIS with KARE data, as well as for the validation analyses using GENIE and ARIRANG, the QC of SNPs and subjects was conducted using PLINK[22] and oneTOOL[23]. We excluded SNPs with *P* values on the Hardy–Weinberg equilibrium (HWE) analysis $< 10^{-5}$, minor allele frequencies (MAFs) < 0.05, and genotype call rates $< 95\%$. Furthermore, we excluded subjects with missing genotype call rates $> 5\%$ or sex-based inconsistencies. After QC, 311,556 SNPs and 7473 participants with FEV_1/FVC ratio ≥ 70 were included for whole-genome imputation.

Genotype imputation. For GWIS with KARE data, whole-genome imputation was performed using SHAPEIT2 and IMPUTE2 for pre-phasing data and genotype imputations. The 1000 Genomes Phase 3 was used as the reference panel. To maintain imputation quality, the estimated imputation accuracy for imputed SNPs was evaluated using the INFO metric, and any imputed SNPs with INFO < 0.5 were eliminated. The standard QC procedure was also applied for imputed SNPs, and 3,351,033 SNPs from 7473 participants were used for the GWIS discovery study (Fig. 1).

For the validation analyses, genotypes comprising the most significant SNPs were not originally genotyped, and target imputation was conducted. Target imputation for regions containing significantly associated SNPs was performed using IMPUTE2 with a buffer size of 5 million bp for each target SNP.

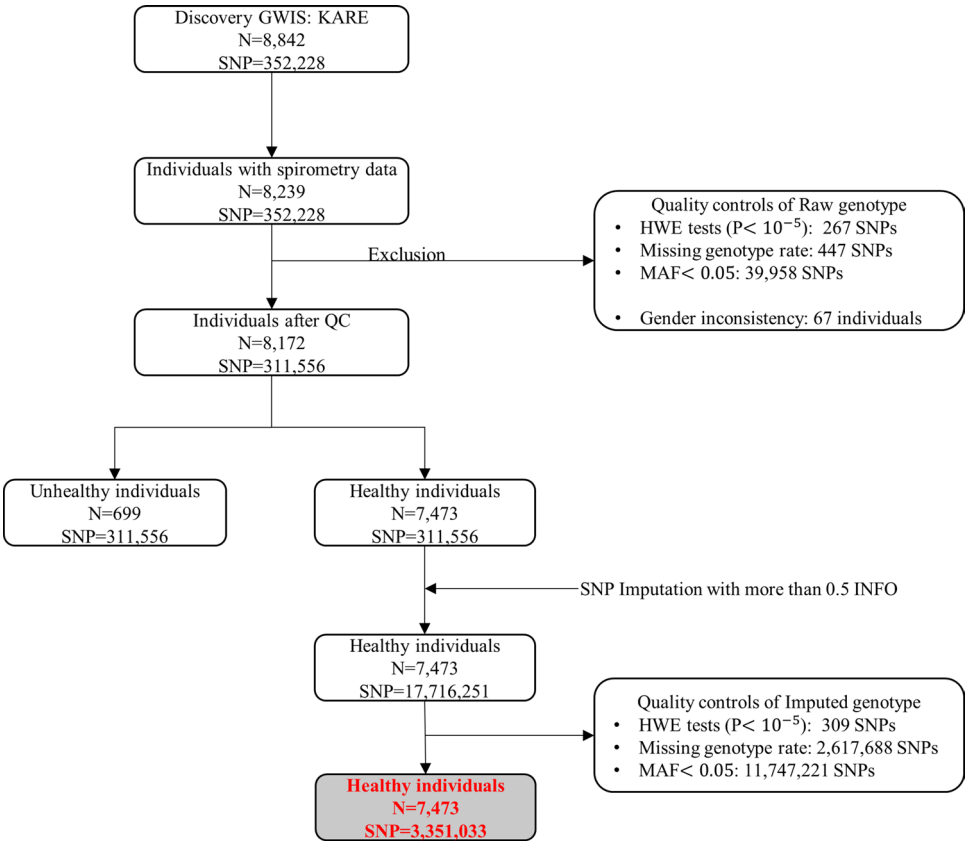

Figure 1. Flow diagram for KARE cohort Fig. 1 explains how the individuals and SNPs were included and excluded. After quality controls and imputations, finally 7473 healthy individuals and 3,351,033 SNPs were used for analyses.

GWIS with KARE data. The GWIS discovery study of the FEV_1/FVC ratios with KARE data was conducted for healthy individuals. We found that the most of our samples do not have any COPD, and the number of COPD patients is only 699. Genetic association analysis needs large sample sizes[24], and thus we decided to focus on healthy individuals with FEV_1/FVC ratio ≥ 70. To handle heteroscedasticity in pulmonary function between non-smokers and smokers, the weighted least squares regression was used with inverse variance weights according to smoking status (non-smoker or smoker). To assess the appropriateness of the weighted least analysis, we compared its Akaike Information Criterion with the linear regression coefficient, and found that the weighted least squares regression had better fit. Age, sex, BMI, age × sex, and pack-years of smoking were included as covariates. Principal component (PC) scores were estimated from the genetic relationship matrix, and 10 PC scores corresponding to the 10 largest eigenvalues were included as covariates to adjust the population substructure. For the GWIS between each SNP and pack-years of smoking, we fitted the following weighted least squares regression. Considering y_j the FEV_1/FVC values for smoking status j, and j=0 and 1, indicating non-smokers and smokers, respectively,

$$y_j \sim \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 BMI + \beta_4 age \times sex + \beta_5 pack\ years + \sum_{k=1}^{10} \tau_k PC^k + \beta_6 SNP + \beta_7 SNP \times pack\ years + \epsilon_j, where \epsilon_j \sim N(0, w_j\sigma^2).$$

Here, for w_1 and w_2, we estimated the residual variances from the linear regressions with only non-smokers and smokers, respectively, and the inverse of residual variances was used.

To identify SNPs interacting with pack-years of smoking on FEV_1/FVC values, we considered $H_0 : \beta_6 = \beta_7 = 0$, and the hypothesis was tested by the likelihood ratio (LR) test with two degrees of freedom (DF) for healthy individuals. To adjust the multiple testing issue, the P values for testing $H_0 : \beta_6 = \beta_7 = 0$ was set to the genome-wide significant level of 5×10^{-8}.

Validation studies. The genome-wide significant SNPs resulting from the GWIS discovery study were replicated using the GENIE and ARIRANG datasets. For the ARIRANG dataset, similar to that performed for the GWIS discovery analysis, healthy individuals were included and the weighted least squares regression was fitted. However, for GENIE, the former smokers were excluded because they consisted of participants who had

regular health check-ups and consults for health improvement, including smoking cessation, regular exercise, etc.[19] Even a short interference of 3 min is said to significantly increase the rate of smoking cessation among smokers[25], and this could bias the data. For the weighted least squares regression of the GENIE data, the weight was estimated using non-smokers and smokers.

A P value < 0.05 was set as the significance level in all analyses.

Topologically associating domains, PrediXcan, and expression quantitative trait loci analyses. Topologically Associating Domains (TADs) are genomic regions that exhibit high levels of chromatin interactions within a region or domain, but with little or no interaction with external regions[26]. These domains are consistent across cell types and highly conserved across species, indicating that the TADs properties are strongly conserved in mammals[27]. TADs were considered to identify boundaries where causal variants can greatly influence tissue-independent function[28]. We used the web-based 3D Genome Browser[29] to identify TADs of significant SNPs from the GWIS and to confirm interactive protein-coding genes within TADs. We selected human tissues per the hg19 assembly and explored available high-throughput chromosome conformation capture (Hi-C) data from lung tissue obtained from donor 1 (Accessing number: SRX2179252 from GEO database).

PrediXcan, a gene-based approach for identifying genes associated with the phenotype of interest[30], imputes unobserved gene expression levels from genotypes and analyzes associations between imputed gene expression and phenotype. The imputation model for gene expression was developed for 48 different human tissues with Genotype-Tissue Expression (GTEx) V7 data. PrediXcan was used to impute gene expression of lung tissue and its association with the FEV_1/FVC ratio.

We analyzed expression quantitative trait loci (eQTL) to investigate genetic variants associated with gene expression levels. For the eQTL analysis, we used the GTEx portal providing reference resources of genetic variation and gene regulation in diverse human tissues[31].

SNP-exposure independence. SNP-by-environment interactions can be significant in the absence of true SNP-by-environment interactions in cases of SNP-environment dependencies[32]. Thus, correlations between most genome-wide significant SNPs from the GWIS discovery study and smoking were assessed. As smoking variables, we considered smoking status and pack-years, which were considered independent responses. SNPs were considered covariates for both scenarios. Smoking status was either non-smoker or smoker, and logistic regression analysis was conducted. For pack-years, linear regressions were performed.

Research ethics approval. This study complies with the scholarly and ethical conduct in research involving human participants. All study participants provided informed consent, and the study design was approved by the Institutional Review Board (IRB) at Seoul National University (IRB No. E1605/E002-003). All methods were performed in accordance with the relevant guidelines and regulations.

Results

GWIS of the FEV_1/FVC ratio among the healthy individuals from the KARE cohort. The GWIS of the FEV_1/FVC ratio was conducted for the 7473 healthy individuals and 3,351,033 SNPs that passed the QC (Fig. 1). The clinical characteristics of the healthy individuals in the KARE cohort are shown in Table 1. The quantile–quantile plot in Fig. 2A shows that GWIS statistics retained the nominal significance level (variance inflation factor = 1.002). The Manhattan plot in Fig. 2B shows that 11 SNPs at 6p21 reached genome-wide significance levels. As shown in Supplementary Fig. 1, these 11 SNPs were distributed in two separate linkage disequilibrium (LD) block. The regional plot in Fig. 3 shows that the most significant SNPs were found near *TNXB*, with many other proximal genes. The most statistically significant result was obtained for rs10947231, an intronic SNP located in *TNXB*, with $P_{LR} = 2.23 \times 10^{-12}$ and corresponding $P_{SNP} = 1.42 \times 10^{-10}$ and $P_{INT} = 0.84$ (Table 2). Here, P_{LR} indicates the likelihood ratio test for $H_0 : \beta_{SNP} = \beta_{SNP*pack-years} = 0$ vs $H_1 : not H_0$.

The second genome-wide significant region was found near the *NOTCH4* intron. In the linkage disequilibrium (LD) block, the rs8192575 SNP showed a genome-wide significant overall effect. The P_{LR} was 1.18×10^{-8} (Table 2). The coefficients for the SNP and interaction effects were 0.821 and − 0.02, respectively ($P_{SNP} = 1.77 \times 10^{-9}$ and $P_{INT} = 0.0165$). These results indicated that if the genotype of rs8192575 is GG, the FEV_1/FVC ratio tends to increase to approximately 0.821×2; however, for smokers, the FEV_1/FVC ratios decrease to approximately -0.02×2 per pack-year. Figure 4 indicates a significant difference in the FEV_1/FVC ratios between non-smokers and smokers. Figure 5 presents box-plots in accordance with the smoking status, age, and genotypes of rs8192575. In healthy individuals, the FEV_1/FVC ratios are consistently larger for non-smokers with genotypes GG and GC than for non-smokers with CC genotypes; however, for smokers, there are no significant differences in FEV_1/FVC by allele G in rs8192575. Estimated FEV_1/FVC ratios in accordance with the pack-years is shown in Fig. 6. In Fig. 6, the decreasing rate of FEV_1/FVC is greater for individuals with genotypes GG and GC than genotypes CC. This indicates that the effect of allele G of rs8192575 is modified by pack-years of smoking.

Other nine significant SNPs were belonging in the same LD blocks with rs10947231 and rs8192575 (Supplementary Fig. 1), and the results are summarized in Supplementary Table 1.

Validation analyses. The associations between rs10947231 and rs8192575 and the FEV_1/FVC ratio were further assessed using the healthy individuals in GENIE and ARIRANG cohorts. Clinical characteristics of healthy individuals these datasets are presented in Table 1. Table 2 shows that rs10947231 overall effects were replicated in the GENIE and ARIRANG data at $P < 0.1$ ($P_{LR} = 0.0698$ for GENIE; $P_{LR} = 0.0714$ for ARIRANG). However the main SNP effects were not significant for GENIE data. Interestingly, for ARIRANG data, rs10947231 showed significant effects on both the SNP and interaction with $P_{SNP} = 0.0415$ and $P_{INT} = 0.0583$.

	KARE	GENIE	ARIRANG
Participants	7473	4413	513
Age (years)	52.0 ± 8.9	49.7 ± 6.8	59.0 ± 7.4
Gender			
Male	3310 (44.3%)	2507 (56.8%)	208 (40.5%)
Female	4163 (55.7%)	1906 (43.2%)	305 (59.5%)
Body mass index (kg/m^2)	24.7 ± 3.1	23.3 ± 2.8	25.3 ± 3.2
Height (cm)	159.8 ± 8.7	165.6 ± 7.9	157.5 ± 8.5
Smoking status			
Non-smokers	4768 (63.8%)	2520 (57.1%)	369 (71.9%)
Former smokers	1140 (15.3%)	1380 (31.3%)	65 (12.7%)
Current smokers	1565 (20.9%)	513 (11.6%)	79 (15.4%)
Pack-years of smoking	22.8 ± 17.2	16.2 ± 24.1	27.5 ± 15.3
FEV_1 (liters)	2.9 ± 0.7	3.0 ± 0.6	2.4 ± 0.6
FVC (liters)	3.6 ± 0.9	3.7 ± 0.8	3.1 ± 0.8
FEV_1/FVC ratio	81.4 ± 5.4	81.7 ± 5.5	79.2 ± 5.1

Table 1. Descriptive statistics of healthy individuals Means of variables and their standard errors are calculated for continuous variables.

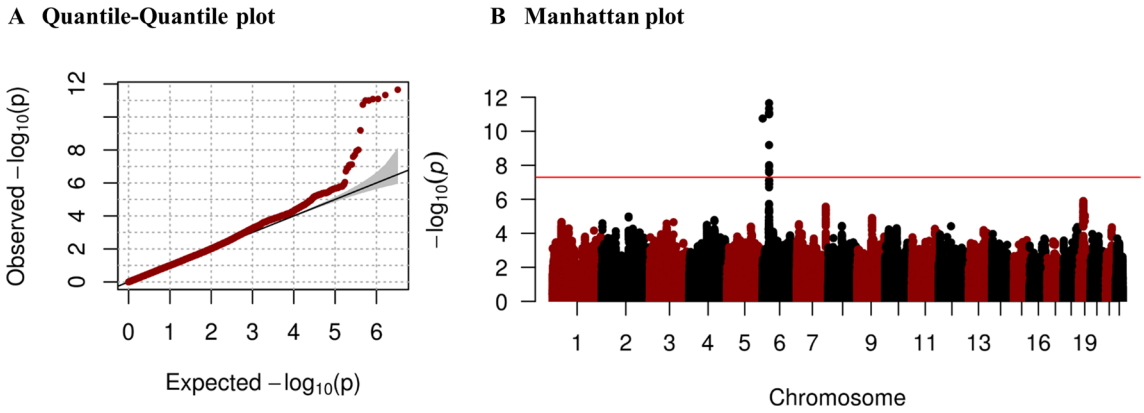

Figure 2. Quantile–quantile plot and Manhattan plot from GWIS with healthy individuals in KARE data (**A**) is obtained from the comparison between observed *P* values quantiles relative and expected quantiles under the uniform distribution (null hypothesis). The variance inflation factor (VIF) was 1.002, suggesting our results are free of systematic *P* value inflation. (**B**) was plotted from the logarithms of the *P* values of 3,351,033 SNPs against its physical chromosomal position. The red line represents genome-wide significance level (5×10^{-8}), and several SNPs located at 6p21 meet this significance level. The plot was generated by software R version 3.6.1 (R Foundation for Statistical Computing; Vienna, Austria).

According to Table 2, the overall effect of rs8192575 was significant for GENIE and ARIRANG data (P_{LR} for GENIE = 0.0173 and P_{LR} for ARIRANG = 0.0457). For the GENIE data, both the main SNP and interaction effects were replicated at 0.05 significance levels ($P_{SNP} = 0.0113$ and $P_{INT} = 0.0454$), and their regression coefficients were 0.533 and -0.031, respectively. For the ARIRANG data, the interaction effect was replicated with -0.067 regression coefficients and 0.0131 *P* value. However, the main SNP effect was not significant ($P_{SNP} = 0.1471$). The results for the other nine significant SNPs are presented in Supplementary Table 2, and are similar to those found for rs10947231 and rs8192575.

TAD, PrediXcan, and eQTL analyses. We considered chromatin TADs containing rs10947231 and rs8192575. TADs were defined using Hi-C to identify chromatin regions with physical contact. Most SNPs associated with human disease or other phenotypes may develop associations through interactions with regulatory elements of a coding gene within the SNP-bearing TAD[33]. Chromatin TAD analysis of lung tissue revealed that *TNXB*, *NOTCH4*, *AGER*, and *C4B* were in the same region (Supplementary Fig. 2). The most active interaction was observed in *NOTCH4*, followed by *AGER*, *TNXB*, and *C4B*, showing no significant differences.

PrediXcan analysis predicted susceptibility on the expression of chromosome 6 genes in lung tissue that regulate the FEV_1/FVC ratio. The results, summarized in Table 3, evidenced *AGER* (*P* value = 8.59×10^{-6}) as the gene most associated with the FEV_1/FVC ratio. These results were significant at the conservative Bonferroni-adjusted 0.05 significance level (*P* value < 1.1×10^{-4}).

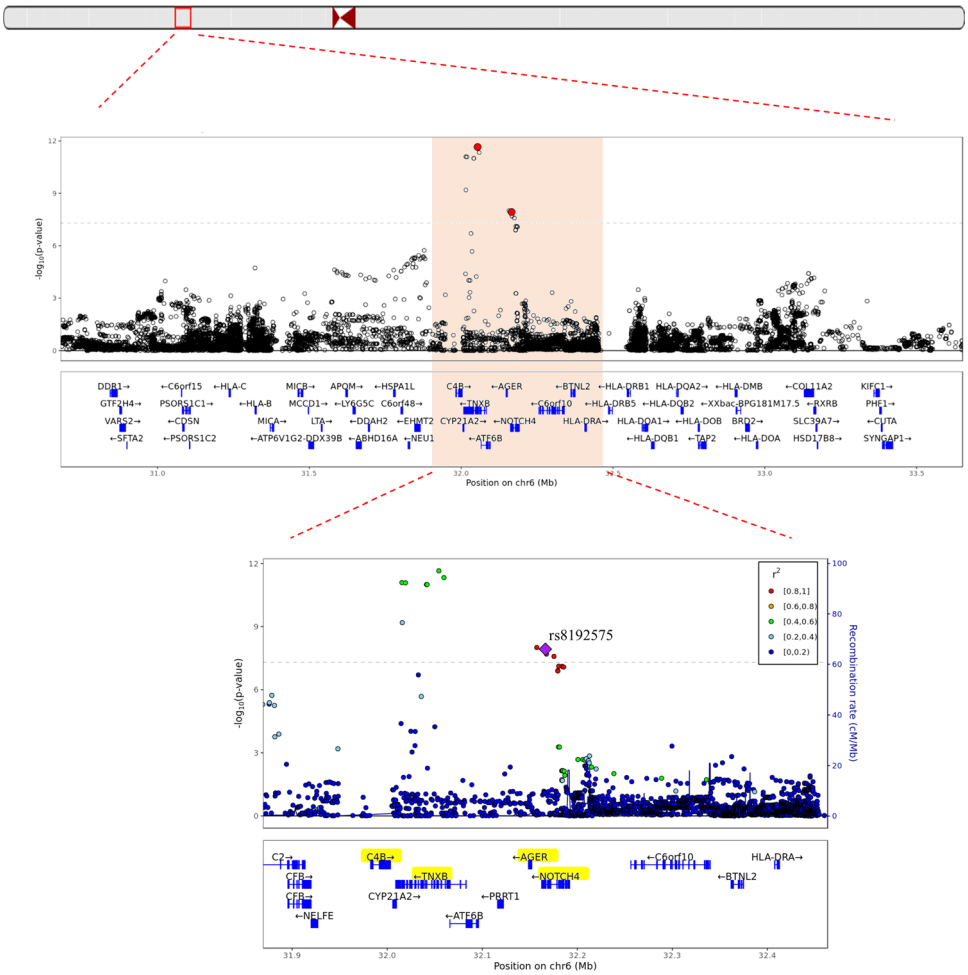

Figure 3. Genomic region on chromosome 6 near rs10947231 and rs8192575 Fig. 3 suggests that the rs10947231 and rs8192575 located at gene dense region. The plot was generated by software R version 3.6.1 (R Foundation for Statistical Computing; Vienna, Austria).

	Data	Minor/Major alleles	MAF	*P* value for HWE	Missing rate	INFO†	β_{SNP}(SE)	P_{SNP}	β_{INT}(SE)	P_{INT}	Overall effects (P_{LR})
rs10947231 Chromosome (BP) 6(32,054,346)											
Discovery	KARE	A/C	0.165	0.767	0.015	0.983	0.835 (0.13)	1.42×10^{-10}	0.001 (0.007)	0.8417	2.23×10^{-12}
Replication	GENIE	A/C	0.18	0.902	0	0.999	0.178 (0.198)	0.3678	0.026 (0.014)	0.0684	0.0698
	ARIRANG	A/C	0.17	0.638	0.016	0.985	0.967 (0,473)	0.04151	− 0.056 (0.03)	0.0583	0.0714
rs8192575 Chromosome (BP) 6(32,166,384)											
Discovery	KARE	G/C	0.15	0.549	0.027	0.976	0.821 (0.136)	1.77×10^{-9}	− 0.02 (0.008)	0.0165	1.18×10^{-8}
Replication	GENIE	G/C	0.163	0.425	0	1	0.533 (0.21)	0.0113	− 0.031 (0.016)	0.0454	0.0173
	ARIRANG	G/C	0.176	0.357	0.006	0.996	0.702 (0.484)	0.1471	− 0.067 (0.027)	0.0131	0.0457

Table 2. Results for rs10947231 and rs8192575 from discovery GWIS For FEV_1/FVC, GWIS was performed on healthy individuals with 3,351,033 SNPs, and the genome-wide significant result are summarized. β_{SNP} and β_{INT} are the coefficients for the main SNP and interaction effects between SNP and pack-years of smoking, respectively. Overall effects indicate *P* values (P_{LR}) for testing the null hypotheses $H_0 : \beta_{SNP} = \beta_{SNP-pack\ years} = 0$ by F test. *BP* physical position (Based on hg19), *MAF* minor allele frequencies, *HWE* Hardy–Weinberg equilibrium, *SE* standard error. †INFO is the imputation quality metric obtained from IMPUTE2.

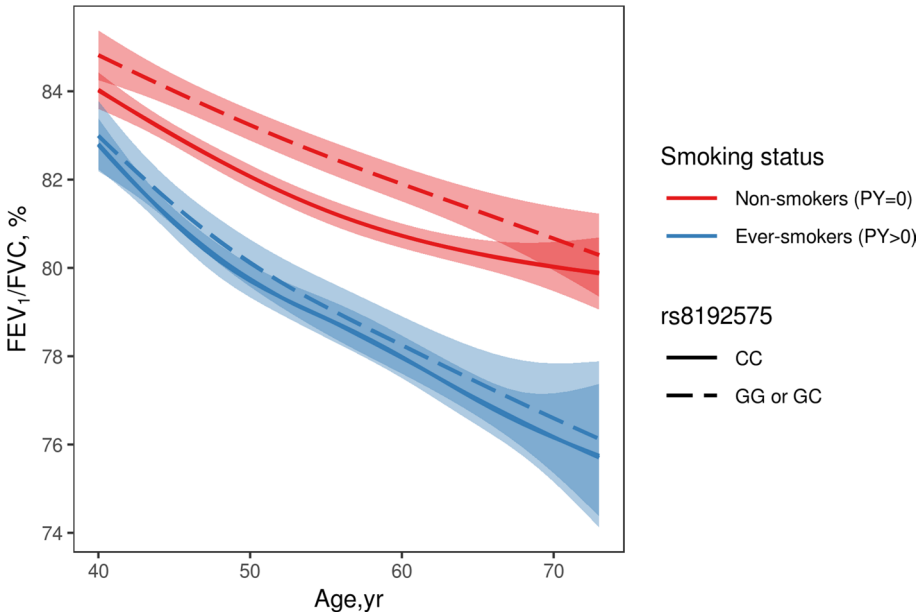

Figure 4. Declines of FEV_1/FVC along age Changes in FEV_1/FVC along age according to smoking status and rs8192575 were plotted by generalized additive models (GAM). Figure 4 suggests that smoking has a significant effect on FEV_1/FVC ratios for healthy individuals. The plot was generated by software R version 3.6.1 (R Foundation for Statistical Computing; Vienna, Austria).

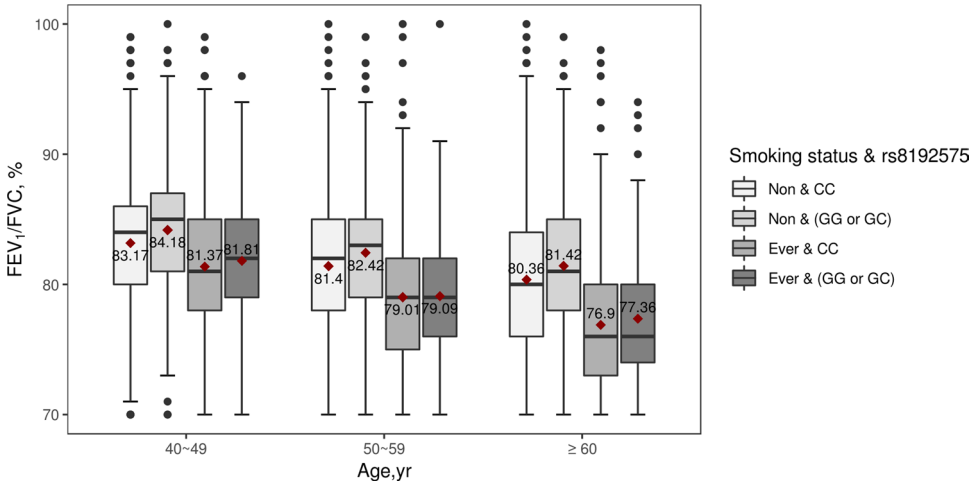

Figure 5. The boxplot of the FEV1/FVC grouped by rs8192575, smoking status, and age from KARE data We visualized the effects of rs819275 on FEV_1/FVC according to smoking status and age. The plot represents the effects of rs8192575 for healthy individuals, and the allele G of rs8192575 has positive effects if healthy individuals have not smoked. The red diamond symbols represent mean value, which is displayed as a number. The plot was generated by software R version 3.6.1 (R Foundation for Statistical Computing; Vienna, Austria).

The eQTL of rs10947231 and rs8192575 were analyzed using GTEx (Supplementary Table 3). For rs10947231, no significant association between rs10947231 genotype and gene expression was found. For rs8192575, many eQTL genes, such as *NOTCH4*, *C4B*, and *AGER* were identified. Interestingly, *AGER* and *C4B* were differentially expressed in lung tissue, based on rs8192575. Thus, we further analyzed *AGER* and *C4B* expression using GTEx V7 data (Supplementary Fig. 3), and this revealed that *AGER* was upregulated in lung tissue, while *C4B* was expressed in adrenal glands and liver tissue. The numbers of transcripts per million kilobases for *C4B* and *AGER* in lung tissue were 5.53 and 1093.06, respectively.

In summary, it is unclear which gene, *AGER* or *C4B*, contributes to the significant effect of rs8192575 on the FEV_1/FVC ratio. However, *AGER* might be a more promising candidate gene due to its higher expression levels in lung tissue compared to that of *C4B*.

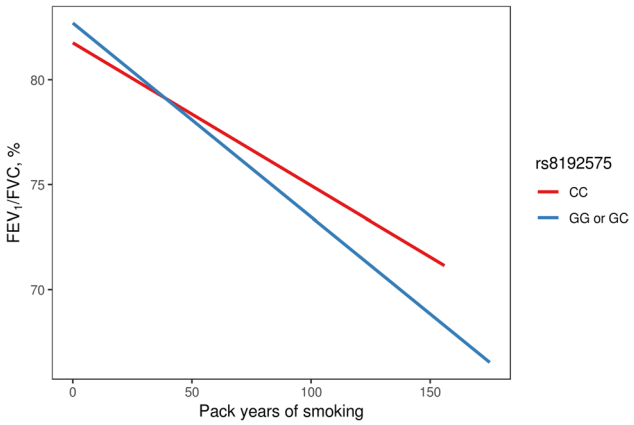

Figure 6. Estimated FEV1/FVC according to pack-years of smoking and rs8192575 The estimated FEV_1/FVC ratio is generated by gene-by-smoking analysis using healthy individuals of KARE. Figure 6 implies that minor alleles G has negative impact in proportion to each increment of pack-years of smoking. The plot was generated by software R version 3.6.1 (R Foundation for Statistical Computing; Vienna, Austria).

Gene	PrediXcan Beta	PrediXcan *P* value
AGER	− 2.797	8.59×10^{-6}
HLA-S	− 0.833	4.60×10^{-5}
C4B	− 0.736	0.0003
MDGA1-201	1.312	0.0012
HLA-J	1.065	0.0019
NKAPL	− 0.941	0.0042
HLA-DPA1	− 0.791	0.0052
HCG4B	0.716	0.0088
TPT1P4	1.522	0.0094
HLA-C	-0.549	0.0099

Table 3. Result of PrediXcan for FEV_1/FVC ratio on KARE using lung tissue prediction model The results for PrediXcan for the top 10 genes on chromosome 6 are summarized.

Independence between rs8192575 and smoking. Significant SNP-by-environment interactions are detected in the absence of true SNP-by-environment interactions if SNP and the environment are correlated. Thus, we investigated whether rs8192575 and smoking are independent, considering smoking status and pack-years as the smoking variables (Table 4). Supplementary Fig. 4 shows a boxplot of pack-years based on rs8192575; this SNP was not significantly associated with pack-years of smoking. Therefore, the statistical significance of the rs8192575-by-smoking interaction may indicate biological interactions, especially among subjects with normal lung function, and this could be true causal effects.

Discussion

The present GWIS including healthy individuals from the KARE dataset found that rs10947231 and rs8192575, located on chromosome 6 (6p21), were significantly associated with the FEV_1/FVC ratio in healthy individuals. Furthermore, rs8192575 showed significant interaction effects with the smoking variable pack-years of smoking. These associations was further investigated using data from two other Korean cohorts (GENIE and ARIRANG). For rs10947231 and rs8192575, both cohorts showed significant overall effects ($P < 0.1$) and significant interaction effects with pack-years. The most significant findings were found in the introns of *TNXB* (rs10947231) and *NOTCH4* (rs8192575). Region 6p21 is gene-dense, including genes *TNXB*, *PPT2*, *C4B*, *NOTCH4*, and *AGER*, among others. We conducted TAD, PrediXcan, and eQTL analyses to confirm the strongly associated genes. We confirmed that rs8192575 was strongly associated with *AGER* in lung tissues, and *AGER* has been previously described as susceptible for lung function and COPD. The rs10947231 does not have any eQTL genes. Its MAF in Europeans were 0.058 and 0.002 with 1000 Genome and gnomAD[34], respectively, and its low MAF may induce non-significance of its eQTL analysis. Considering that rs10947231 is in the TADs block of *AGER* region, and DNA sequences within TAD interact more frequently with each other than those outside TAD, there are some possibilities of significant association between *AGER* and rs10947231.

AGER is a protein coding gene which encodes RAGE belonging to the immunoglobulin superfamily and cell-surface receptor[35]. The RAGE has been extensively studied, and it significantly contributes to lung development and to maintain adult lung homeostasis, as evidenced by its upregulation on the membrane and cytoplasm of both

DATA	*P* value from healthy individuals
Response variable: Pack-years of smoking	
KARE	0.659
GENIE	0.3294
ARIRANG	0.8955
Response variable: Smoking status (never- and ever-smokers for KARE, ARIRANG; never- and current-smokers for GENIE)	
KARE	0.7734
GENIE	0.2881
ARIRANG	0.7462

Table 4. Independence test between rs8192575 and smoking variables We tested gene-smoking dependency. The association between rs8192575 and pack-years of smoking are analyzed by linear regression, and the association between rs8192575 and smoking status are analyzed by logistic regression. For KARE and ARIRANG data the smoking status indicates non- and ever-smokers. For GENIE data smoking status indicates non- and current smokers. *P* value greater than 0.05 suggests that there was no dependency between rs8192575 and smoking variables.

Type 1 alveolar cells and macrophages[36]. A recent study suggested that RAGE upregulation during lung development inhibits alveolar morphogenesis and induces significant changes in morphometric parameters, including a reduction in airspace and an increase in alveolar duct size[37]. Lee et al.[38] reported that the blockade of RAGE is significantly associated with decreased pulmonary inflammation and inhibits the activation of damage-associated molecular patterns in mice exposed to tobacco smoke. Indeed, RAGE expression increased after exposure to tobacco smoke and in COPD patients[39] suggesting that RAGE suppression protects against COPD. The eQTL analysis revealed that the minor allele G of rs8192575 was associated with lower *AGER* expression in lung tissue; thus, *AGER* may be downregulated in individuals with a larger number of G alleles. Figures 4 and 5 show that participants with G alleles also tended to have a greater FEV_1/FVC ratio, suggesting that *AGER* is associated with rs8192575 (This result is consistent with the previous report that RAGE suppression protect against COPD). The circulating soluble RAGE (sRAGE), acting as decoy receptors, has been robustly demonstrated that low sRAGE levels are associated with advanced COPD and lung function decline, which may be counterintuitive with our results. One possible explanation is that the effect of genetic variants on sRAGE protein levels is affected by environmental exposure, which indicates gene × smoking interaction effects. Another possibility is that genetic variants that promote the association between sRAGE and lung disease susceptibility have different mechanism[40].

Nonetheless, this study has some limitations. First, the 6p21 gene-dense region and *C4B* might have affected the eQTL and PrediXcan results. Gene *C4B* is a product of complement C4 activated in the early stage of the mannose-binding lectin pathway[41], and some studies have reported that *C4B* is associated with tissue damage in pulmonary tuberculosis patients[42,43]. However, *C4B* was not upregulated in lung tissues (Supplementary Fig. 3), and its role in pulmonary function is still lesser known than that of *AGER*. Thus, we concluded *AGER* is a more promising candidate gene for COPD. Second, the most significant SNP, rs10947231, showed $P = 1.42 \times 10^{-10}$ for its main effects in the KARE dataset, but only the ARIRANG data replicated this significance. Because rs10947231 was not directly associated with *AGER*, allelic and locus heterogeneities could be possible reasons for the failure to replicate such significant effects in the GENIE dataset. Third, the rs2070600 located in *AGER*, previously reported to be associated with the FEV_1/FVC ratio in the European population and associated with COPD, emphysema, and sRAGE levels, was not examined in our study. This SNP was excluded in our discovery GWIS during genotype QC and it did not show statistically significant interaction effects ($P_{LR} = 1.673 \times 10^{-15}$, $P_{SNP} = 1.87 \times 10^{-13}$, $P_{INT} = 0.984$), potentially due to differences in the study population. These differences may be attributable to differences in genetic ancestry and LD structure among populations. Fourth, the overall and main SNP effects of rs8192575 were significant throughout the genome (significance level $= 5 \times 10^{-8}$). However, SNP-by-smoking interactions were significant at a relatively high significance level, i.e., at 0.05. Analyses of gene-by-environment interactions often present numerous false-negative results, concurrent with the present findings. Hence, a larger sample is necessary to obtain genome-wide significant results to analyze gene-by-environment interactions. Fifth, our functional analyses (eQTL, TAD, and PrediXcan) have some limitations. For eQTL and TAD, the relevance of their results depends on tissue type. COPD and FEV_1/FVC are particularly related to lung tissue and it was chosen. However it is still possible that the other tissue can be a better choice[44]. Furthermore, the prediction model of PrediXcan was built by European but it was applied to Korean. In such case, the prediction accuracy can become worse[45,46].

In conclusion, we identified genome-wide significant effects at the 6p21 region using the FEV_1/FVC of healthy individuals, and rs8192575 showed significant interaction effects with smoking. Indeed, 6p21 is a gene-dense region, as characterized by previous GWAS.[7,8,47] However, significant results were obtained with healthy individuals, and evidence regarding its significant interaction effects with smoking were found in Korean cohort data. The MAF of rs8192575 for the European population was 0.083 with 1000G European samples. Therefore, the lack of significant results so far might be due to low allele frequencies. However, for Koreans, the MAF was relatively high (0.15 to 0.18), and rs8192575 seems to have a significant effect on the Korean population. We expect that these results potentially provide insights into COPD pathogenesis and on the effect of smoking on lung function, concurrent with previous GWAS and biological reports.

Data avilability

Genotype and clinical data for KARE and ARIRANG can be downloaded from https://koreabiobank.re.kr followed by an approval process from Korean NIH. For more detail information for approval process, please contact to biobank@korea.kr. All type data for GENIE cohort is available on request at https://en-healthcare.snuh.org/HPEACEstudy. In addition, all the data analyzed in this article was utilized in previously published articles (KARE: Cho et al. [18], ARIRANG: Huh et al.[21], GENIE: Lee et al.[19].

Received: 13 January 2020; Accepted: 15 July 2020
Published online: 04 August 2020

References

1. Vestbo, J. *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am. J. Respir. Crit. Care Med.* **187**, 347–365 (2013).
2. Hogg, J. C. Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. *The Lancet* **364**, 709–721 (2004).
3. Snider, G. L. Chronic obstructive pulmonary disease: risk factors, pathophysiology and pathogenesis. *Annu. Rev. Med.* **40**, 411–429 (1989).
4. Terzikhan, N. *et al.* Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study. *Eur. J. Epidemiol.* **31**, 785–792 (2016).
5. DeMeo, D. & Silverman, E. α1-Antitrypsin deficiency· 2: Genetic aspects of α1-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk. *Thorax* **59**, 259–264 (2004).
6. Hersh, C. P., DeMeo, D. L. & Silverman, E. K. National Emphysema Treatment Trial state of the art: genetics of emphysema. *Proc. Am. Thoracic Soc.* **5**, 486–493 (2008).
7. Repapi, E. *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat. Genet.* **42**, 36 (2010).
8. Hancock, D. B. *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42**, 45 (2010).
9. Zhou, J. J. *et al.* Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am. J. Respir. Crit. Care Med.* **188**, 941–947 (2013).
10. Kim, W. J. & Do Lee, S. Candidate genes for COPD: current evidence and research. *Int. J. Chronic Obstr. Pulm. Dis.* **10**, 2249 (2015).
11. Silverman, E. K. *et al.* Family study of α1-antitrypsin deficiency: Effects of cigarette smoking, measured genotype, and their interaction on pulmonary function and biochemical traits. *Genet. Epidemiol.* **9**, 317–331 (1992).
12. Aschard, H. *et al.* Evidence for large-scale gene-by-smoking interaction effects on pulmonary function. *Int. J. Epidemiol.* **46**, 894–904 (2017).
13. Hancock, D. B. *et al.* Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS Genet.* **8**, e1003098 (2012).
14. Park, B. *et al.* Genome-wide assessment of gene-by-smoking interactions in COPD. *Sci. Rep.* **8**, 9319 (2018).
15. Castaldi, P. J. *et al.* Impact of non-linear smoking effects on the identification of gene-by-smoking interactions in COPD genetics studies. *Thorax* **66**, 903–909 (2011).
16. Lundbäck, B. *et al.* Not 15 but 50% of smokers develop COPD?—report from the obstructive lung disease in Northern Sweden studies. *Respir. Med.* **97**, 115–122 (2003).
17. Camargo, C. A. Jr. *et al.* Promotion of lung health: NHLBI workshop on the primary prevention of chronic lung diseases. *Ann. Am. Thoracic Soc.* **11**, S125–S138 (2014).
18. Cho, Y. S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* **41**, 527 (2009).
19. Lee, C. *et al.* Health and Prevention Enhancement (H-PEACE): a retrospective, population-based cohort study conducted at the Seoul National University Hospital Gangnam Center Korea. *BMJ Open* **8**, e019327 (2018).
20. Park, B. *et al.* Genetic Polymorphisms Associated with the Neutrophil-Lymphocyte Ratio and Their Clinical Implications for Metabolic Risk Factors. *J. Clin. Med.* **7**, 204 (2018).
21. Huh, J. H. *et al.* A prospective study of fatty liver index and incident hypertension: the KoGES-ARIRANG Study. *PLoS ONE* **10**, e0143560 (2015).
22. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
23. Song, Y. E. *et al.* ONETOOL for the analysis of family-based big data. *Bioinformatics* **1**, 3 (2018).
24. Liu, C.-Y., Maity, A., Lin, X., Wright, R. O. & Christiani, D. C. Design and analysis issues in gene and environment studies. *Environ. Health* **11**, 93 (2012).
25. Tobacco, T. C. P. G. T. A clinical practice guideline for treating tobacco use and dependence: 2008 update: a US public health service report. *Am. J. Prev. Med.* **35**, 158–176 (2008).
26. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).
27. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
28. Way, G. P., Youngstrom, D. W., Hankenson, K. D., Greene, C. S. & Grant, S. F. Implicating candidate genes at GWAS signals by leveraging topologically associating domains. *Eur. J. Hum. Genet.* **25**, 1286 (2017).
29. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
30. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091 (2015).
31. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580 (2013).
32. Dudbridge, F. & Fletcher, O. Gene-environment dependence creates spurious gene-environment interaction. *Am. J. Hum. Genet.* **95**, 301–307 (2014).
33. Grubert, F. *et al.* Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).
34. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
35. Neeper, M. *et al.* Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins. *J. Biol. Chem.* **267**, 14998–15004 (1992).
36. Sorci, G., Riuzzi, F., Giambanco, I. & Donato, R. RAGE in tissue homeostasis, repair and regeneration. *Biochim. Biophys. Acta (BBA)-Mol. Cell Res.* **1833**, 101–109 (2013).

37. Fineschi, S. *et al.* Receptor for advanced glycation end products contributes to postnatal pulmonary development and adult lung maintenance program in mice. *Am. J. Respir. Cell Mol. Biol.* **48**, 164–171 (2013).
38. Lee, H. *et al.* Blockade of RAGE ameliorates elastase-induced emphysema development and progression via RAGE-DAMP signaling. *FASEB J.* **31**, 2076–2089 (2017).
39. Wu, L., Ma, L., Nicholson, L. F. & Black, P. N. Advanced glycation end products and its receptor (RAGE) are increased in patients with COPD. *Respir. Med.* **105**, 329–336 (2011).
40. Yonchuk, J. G. *et al.* Circulating soluble receptor for advanced glycation end products (sRAGE) as a biomarker of emphysema and the RAGE axis in the lung. *Am. J. Respir. Crit. Care Med.* **192**, 785–792 (2015).
41. Lachmann, P. J. Microbial immunology: a new mechanism for immune subversion. *Curr. Biol.* **8**, R99–R101 (1998).
42. Wang, C. *et al.* Serum complement C4b, fibronectin, and prolidase are associated with the pathological changes of pulmonary tuberculosis. *BMC Infect. Dis.* **14**, 52 (2014).
43. Jiang, T.-T. *et al.* Serum amyloid A, protein Z, and C4b-binding protein β chain as new potential biomarkers for pulmonary tuberculosis. *PLoS ONE* **12**, e0173304 (2017).
44. Brandsma, C.-A. *et al.* Lung ageing and COPD: is there a role for ageing in abnormal tissue repair?. *Eur. Respir. Rev.* **26**, 170073 (2017).
45. Mikhaylova, A. V. & Thornton, T. A. Accuracy of gene expression prediction from genotype data with PrediXcan varies across and within continental populations. *Front. Genet.* **10**, 261 (2019).
46. Li, B. *et al.* Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. *Pac. Symp. Biocomput.* **23**, 448–459 (2018).
47. Kim, W. J. *et al.* Genome-wide association studies identify locus on 6p21 influencing lung function in the K orean population. *Respirology* **19**, 360–368 (2014).

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI15C2165), and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1A09919610).

Author contributions

B.P. and S.W. wrote the main manuscript text, and B.P., J.A., and W.K. analyzed and interpreted the results. H.Y.K., S.B.K., B.O., K.J.J., and S.H.J. provide the data. W.J.K., M.H.C., E.K.S., T.P., and S.W. revised this paper critically for important intellectual content. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-70092-0.

Correspondence and requests for materials should be addressed to T.P. or S.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.